Langdon, R., Richmond, R., Hemani, G., Zheng, J., Wade, K., Carreras-Torres, R., Johansson, M., Brennan, P., Wootton, R., Munafo, M., Davey Smith, G., Relton, C., Vincent, E., Martin, R., & Haycock, P. (2019). A phenome-wide Mendelian randomization study of pancreatic cancer using summary genetic data. *Cancer Epidemiology, Biomarkers and Prevention*. https://doi.org/10.1158/1055-9965.EPI-19-0036

Peer reviewed version

Link to published version (if available):
10.1158/1055-9965.EPI-19-0036

A phenome-wide Mendelian randomization study of pancreatic cancer using summary genetic data

Ryan J Langdon[1,2], BSc; Rebecca Richmond[1,2], PhD; Gibran Hemani[1,2], PhD; Jie Zheng[1,2], PhD; Kaitlin H Wade[1,2], PhD; Robert Carreras-Torres[3], PhD; Mattias Johansson[4], PhD; Paul Brennan[4], PhD; Robyn Wootton[1,2,5], BSc; Marcus Munafo[1,2,5], MA, MSc, PhD; George Davey Smith[1,2], DSc; Caroline Relton[1,2], PhD, BSc, PGCE; Emma E Vincent[1,2], PhD; Richard M Martin[1,2] *, BMedSci, BMBS, MSc, PhD; Philip C Haycock[1,2] *, PhD

*These authors contributed to the manuscript equally

[1] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK

[2] Bristol Medical School, Department of Population Health Sciences, University of Bristol, Bristol, UK

[3] Biomarkers and Susceptibility Unit, IDIBELL - Bellvitge Biomedical Research Institute, Barcelona, Spain

[4] Section of Genetics, International Agency for Research on Cancer (IARC), Lyon, France

[5] UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, Bristol, UK

Correspondence to: PC Haycock philip.haycock@bristol.ac.uk; Oakfield House, Oakfield Grove, Clifton, BS8 2BN; +44 (0) 117 3310088

Financial Conflict of Interest: The authors have no relevant conflicts of interest to report

Word count: 3819

Abstract

Background: The 5-year mortality rate for pancreatic cancer is amongst the highest of all cancers. Greater understanding of underlying causes could inform population-wide intervention strategies for prevention. Summary genetic data from genome-wide association studies (GWAS) have become available for thousands of phenotypes. These data can be exploited in Mendelian randomization (MR) phenome-wide association studies (PheWAS) to efficiently screen the phenome for potential determinants of disease risk.

Methods: We conducted an MR-PheWAS of pancreatic cancer using 486 phenotypes, proxied by 9124 genetic variants, and summary genetic data from a GWAS of pancreatic cancer (7,110 cancer cases; 7,264 controls). Odds ratios and 95% confidence intervals per 1 SD increase in each phenotype were generated.

Results: We found evidence that previously reported risk factors of body mass index (1.46; 1.20 to 1.78) and hip circumference (1.42; 1.21 to 1.67) were associated with pancreatic cancer. We also found evidence of novel associations with metabolites that have not previously been implicated in pancreatic cancer: fibrinogen-cleavage peptide (1.60; 1.31 to 1.95) and O-sulfo-L-tyrosine (0.58; 0.46 to 0.74). An inverse association was also observed with lung adenocarcinoma (0.63; 0.54 to 0.74).

Conclusions: Markers of adiposity (BMI and hip circumference) are potential intervention targets for pancreatic cancer prevention. Further clarification of the causal relevance of fibrinogen cleavage peptides and O-sulfo-L-tyrosine in pancreatic cancer aetiology is required, as is the basis of our observed association with lung adenocarcinoma.

Impact: For pancreatic cancer, MR-PheWAS can augment existing risk factor knowledge and generate novel hypotheses to investigate.

Introduction

People diagnosed with pancreatic cancer have a very poor prognosis, with a less than 5% five-year survival rate(1); symptoms do not manifest until the cancer is at an advanced stage and the disease is rarely detected early. Greater understanding of the aetiology of pancreatic cancer could reduce its burden by informing whole-population or risk-stratified prevention strategies.

Risk factors previously reported for pancreatic cancer include cigarette smoking(2), type 2 diabetes(3), adiposity(4) and chronic pancreatitis(5). However, these reports are based on observational epidemiological studies, which are prone to unmeasured or residual confounding and reverse causation, precluding robust causal inference. Furthermore, conventional epidemiological studies often test a narrow set of hypotheses using prior subject knowledge, typically based on other observational studies. Whilst essential, these approaches can constrict a field of research, and preoccupation with previously-hypothesised risk factors can prevent both the identification of novel risk factors and prioritization of causal associations(6).

Mendelian randomization (MR) is a well-established type of instrumental variable (IV) analysis that addresses some of the shortcomings of conventional observational studies by using genetic anchors to appraise the causal relevance of exposures in disease(7). It is an increasingly recognised and powerful tool for identifying causes of a broad spectrum of outcomes, including cancer(8, 9). Two-sample MR uses summary-level data from published genome-wide association studies (GWASs) to allow causal appraisal of hypothesized exposure-outcome associations using *gene*-exposure and *gene*-outcome associations collected in separate studies(10-12). This method can be extended to appraise causality in a hypothesis-free manner, appraising 1-to-many, many-to-1 or many-to-many exposure-outcome combinations, in an approach known as a MR phenome-wide association study (MR-PheWAS)(13, 14).

Here, we used MR-PheWAS to screen the phenome for potential causes of pancreatic cancer. Our aims were twofold: to identify potentially novel causes of pancreatic cancer that may not have been captured using previous epidemiological approaches, and to prioritise hypotheses identified in current literature.

Materials and Methods

Data preparation

Genetic instruments for phenotypes

Two-sample MR was conducted using the TwoSampleMR R package(15). Genetic data on cognitive, anthropometric, metabolic, immune and behavioural phenotypes were obtained from the MR-Base database of harmonised GWAS summary data (**Supplementary Figure 1**). All phenotypes possessing robust genetic proxies (defined as $P<5e^{-8}$) with which to conduct MR analyses were considered for further analysis (N=523). Duplicate (N=17) and non-European studies (N=8) were excluded from the analysis at this stage, leaving 498 potential phenotypes for analysis. Genetic instruments for each phenotype were single-nucleotide polymorphisms (SNPs) independently associated with the phenotype of interest after linkage disequilibrium (LD) clumping (window=10,000kb; $r^2=0.1$). For each identified SNP, the reported effect size was expressed as a one standard deviation (SD) increase in the level of the phenotype per risk allele, along with the standard error (SE). In the case of a binary phenotype (e.g. presence or absence of coronary heart disease), the reported effect size was expressed as a log-odds ratio (OR). The single largest or most recent summary GWAS data were used per phenotype, systematically prioritised by the instrument extraction function (*extract_instruments)* of the TwoSampleMR R package and preventing bias from sample overlap from multiple GWAS for exposure phenotypes. For each genetic variant associated with the identified phenotypes, effect-estimates and SEs were extracted from the summary genetic data for pancreatic cancer.

To harmonise the data, effect alleles in the pancreatic cancer summary data were coded to reflect the phenotype increasing allele, using allele frequencies to resolve strand ambiguities for palindromic SNPs (A/T or C/G). Those phenotypes that did not have genetic variants in the pancreatic cancer GWAS were excluded (N = 12), resulting in a final list of 486 phenotypes on which to perform MR analyses. These phenotypes are tabulated in **Supplementary Table 1**, which details the phenotype

name, the corresponding author or contributing consortium, the sample size of the contributing GWAS, the number of SNPs in the GWAS and the original units of each phenotype.

Pancreatic cancer data

GWAS data from people of European descent with pancreatic cancer and matched controls were obtained from the PanScan (12 studies) and PanC4 (10 studies) consortia through the National Centre for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP)(16) (Study Accession: phs000206.v3.p2 and phs000648.v1.p1; project reference #9314). PanScan and PanC4 were initially published in three releases: PanScan I (1,788 cases and 1,769 controls), PanScan II (1,696 cases and 1,563 controls) and PanC4 (3,626 cases and 3,932 controls)(17-19). The samples were originally genotyped using Illumina HumanHap550 (PanScan I), Human610-Quad (PanScan II) and HumanOmniExpressExome-8v1 (PanC4) arrays. A summary of the characteristics of the consortia and contributing studies is provided in **Supplementary Tables 2a** and **2b**.

Initial quality control steps and analyses were performed within each publication set at the International Agency for the Research of Cancer (IARC), Lyon. After removing duplicates, related samples, samples with sex discrepancy and population outliers, 7,110 cases and 7,264 controls remained across the three combined consortia. Genotype imputation was performed using the Michigan Imputation Server(20). Genotypes were pre-phased using SHAPEIT v2(21) and imputed with Minimach v3(22) using the Haplotype Reference Consortium panel(23). After imputation, SNPs with an imputation quality (R^2) lower than 0.7 were removed from the datasets. Effect estimates for pancreatic cancer risk were obtained after adjusting for age, sex and principal components for population stratification using R software (R version 3.3.1). Results from each PanScan release were then combined using a fixed-effects inverse-variance approach implemented in METAL(24). Finally, outcome data were converted from a "chromosome: position" format to reference SNP cluster ID (rsID), using the "biomaRt" R package(25) with human genome build 19 (hg19) as reference, to generate SNP IDs that were in the format expected by the TwoSampleMR R package.

Power Calculations

Low power can be a limitation of MR because genetic polymorphisms typically explain a small amount of phenotypic variance. We calculated *a-priori* power, based on a median sample size of 14905, across a range of pre-defined phenotypic variances and effect sizes (**Figure 1**). The median variance explained by SNP IVs for our 486 phenotypes was 3.3%. At this variance, our power calculations indicated we had 80% power to detect a minimum odds ratio (OR) of 1.14 (beta of 0.13), at an alpha of 0.05.

Mendelian randomization analyses

We used maximum likelihood(26) and multiplicative random effects inverse-variance weighted (MRE IVW)(27, 28) MR analyses when the number of SNPs instrumenting a phenotype was greater than 1. Both have been proposed for MR analyses when using summary genetic data with phenotype instruments containing multiple SNPs(29). An MRE model allows for heterogeneity between the causal estimates targeted by the genetic variants by allowing over-dispersion in the regression model. Under-dispersion is not permitted (in case of under-dispersion, the residual standard error is set to 1, as in a fixed-effect analysis). For phenotypes instrumented by a single SNP, we derived Wald ratio effect estimates(29, 30). Results were expressed ORs with a corresponding 95% confidence interval (CI) per 1 standard deviation (SD) increase in continuous traits (e.g. height), and as ORs with 95% CI per 2-fold increase(31) (interpreted as a doubling in odds) for binary traits (e.g. type 2 diabetes).

To correct for multiple testing, the correlation structure amongst the analysed phenotypes was estimated using PhenoSpD(32), which implements principal component analysis to identify independent variables using GWAS summary-level statistics. Firstly, a correlation matrix of phenotypes was built using metaCCA(33), estimating Pearson pair-wise correlations between the GWAS summary data for each phenotype. Once the correlation matrix was built, the effective number of independent phenotypes was estimated using matrix spectral decomposition(34, 35). PhenoSpD overestimates the number of independent variables as it treats phenotypes from separate studies as entirely independent when it is likely they are not. Therefore, our Bonferroni correction for multiple testing is likely particularly conservative.

Sensitivity analyses

MR-Egger regression(36) was used as a sensitivity analysis to detect bias due to horizontal pleiotropy in the causal estimates. Horizontal pleiotropy is where a genetic variant affects the outcome via a different biological pathway from the phenotype under investigation and is a violation of a key assumption of MR (see **Supplementary Figure 3**). MR-Egger regression performs a weighted linear regression of the SNP-disease and SNP-phenotype associations, the intercept of which is not constrained to the origin and can therefore be used to detect and estimate the magnitude of horizontal pleiotropy(36). Deviation from the origin in an MR-Egger regression may suggest the effect of the SNP is operating via a separate pathway. MR-Egger is less efficient when the number of SNPs is low (N<4), therefore we omitted this analysis where phenotypes were proxied by 3 or fewer SNPs. Additionally, we assessed evidence of heterogeneity between SNPs (another potential indication of horizontal pleiotropy and other violations of MR assumptions) for the causal effect estimates of the phenotype on pancreatic cancer using forest plots and Cochran's Q test. Finally, we investigated whether effect estimates were different in men and women, and across the different studies within each consortium using the Q test for heterogeneity (37).

Results

Using PhenoSpD, we estimated that the 486 phenotypes we investigated corresponded to 312 independent tests(32). To aid interpretation of our MR analyses, we set a P-value threshold of $1.6e^{-4}$ (0.05/312) to suggest evidence of association and to prioritise phenotypes for follow-up analyses. Five phenotypes were associated with pancreatic cancer at this threshold (**Figure 2**, **Table 1**). The results of the MR analyses for all phenotypes are shown in **Supplementary Table 3**. Of the 5 associations, 2 were inversely related to pancreatic cancer: lung adenocarcinoma (OR for pancreatic cancer [95% CI]: 0.63 [0.54 to 0.74] per doubling in the odds of lung adenocarcinoma; P: $1.68e^{-8}$) and the metabolite O-sulfo-L-tyrosine (0.58 [0.46 to 0.74] per SD increase; P: $2.45e^{-5}$). The other 3 phenotypes were positively related to pancreatic cancer (OR [95% CI per SD increase]: ADpSGEGDFXAEGGGVR* (a fibrinogen cleavage peptide) (1.60 [1.31 to 1.95]; P: $1.50e^{-3}$); hip circumference (1.42 [1.21 to 1.67]; P:

3.92e^{-4}); and body mass index (1.46 [1.20 to 1.78]; P: 4.02e^{-6}). Maximum likelihood effect estimates were consistent with IVW estimates for these associations (**Table 1**).

There was evidence that the effect of hip circumference on pancreatic cancer varied by pancreatic cancer consortium (Q: 26.52, P: 1.75e^{-06}), but this was not observed for ADpSGEGDFXAEGGGVR* (Q: 1.57, P: 0.46), lung adenocarcinoma (Q: 0.19, P: 0.91), O-sulfo-L-tyrosine (Q: 4.80, P: 0.09) or BMI (Q: 1.56, P: 0.46) (**Figure 3**). There was also evidence that effects varied by sex for hip circumference (Q: 25.3, P: 4.86e^{-7}), but not ADpSGEGDFXAEGGGVR* (Q: 2.67, P: 0.10), lung adenocarcinoma (Q: 0.43, P: 0.51), O-sulfo-L-tyrosine (Q: 0.13, P: 0.72) or BMI (Q: 0.00, P: 0.95) (**Figure 4**).

There was clear evidence of heterogeneity in associations with pancreatic cancer amongst the individual SNPs used as IVs for body mass index (Q: 186.61, P: 0.01), hip circumference (Q: 105.67, P: 4.02e^{-6}), lung adenocarcinoma (Q: 36.64, P: 5.47e^{-8}), ADpSGEGDFXAEGGGVR* (Q: 10.08, P: 1.50e^{-3}) and O-sulfo-L-tyrosine (Q:13.45, P: 2.45e^{-4}) (**Supplementary Figure 2a-e**). The observed heterogeneity is consistent with violations of IV assumptions, such as the presence of horizontal pleiotropy. Intercept tests from MR-Egger regression did not, however, indicate strong evidence for bias from unbalanced pleiotropy for body mass index (OR: 1.00, 95% CI: 0.98 to 1.02 P: 0.84) and hip circumference (OR: 1.00, 95% CI: 0.98 to 1.03, P: 0.75). In addition, effect estimates from MR-Egger regression for hip circumference (OR: 1.18, 95% CI: 0.54 to 2.50, P: 0.68) and body mass index (OR: 1.35, 95% CI: 0.71 to 2.51, P: 0.36) were broadly compatible with results based on the maximum likelihood and IVW methods, albeit with wide confidence intervals (see **Table 1**). Whilst an inverse association was seen for lung adenocarcinoma and pancreatic cancer, the intercept from MR-Egger regression was negative (OR: 0.83, 95% CI: 0.51 to 1.35 P: 0.52) and the slope was in the opposite direction to the effect observed in the main analysis (OR: 1.57, 95% CI: 0.20 to 11.17, P: 0.71).

ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine were both instrumented by 2 SNPs, thus MR-Egger could not be used to assess horizontal pleiotropy for these phenotypes. Associations for both metabolites appeared to be largely driven by rs651007 (a SNP found in the ABO blood group region). The evidence for a causal effect of ADpSGEGDFXAEGGGVR* on pancreatic cancer was weaker for

the second SNP (rs601338) used to instrument ADpSGEGDFXAEGGGVR* (OR: 1.04, 95% CI: 0.75 to 1.44, P:0.81). Similarly, the evidence for a causal effect of O-sulfo-L-tyrosine was weaker for the other SNP (rs6151429) used to instrument O-sulfo-L-tyrosine (OR: 0.84, 95% CI: 0.62 to 1.14, P:0.26).

Of the most established observational phenotypes with pancreatic cancer (smoking, diabetes, chronic pancreatitis and adiposity)(38, 39) only pancreatitis could not be instrumented and only adiposity passed our P-value threshold for further evaluation. The odds ratio (95% CI) for pancreatic cancer per SD increase in cigarettes smoked per day was 1.27 (0.67 to 2.42, P: 0.46) and was 1.02 (0.95 to 1.10, P: 0.56) per doubling in the odds of type 2 diabetes (**Supplementary Table 3**).

Discussion

We undertook an MR-PheWAS of the association of 486 phenotypes with pancreatic cancer, including cognitive, anthropometric, metabolic, immune and behavioural phenotypes. We provide evidence that 5 of the 486 phenotypes we tested were associated with pancreatic cancer: BMI; hip circumference; ADpSGEGDFXAEGGGVR* (a fibrinogen cleavage peptide); O-sulfo-L-tyrosine; and lung adenocarcinoma.

The association of higher BMI with risk of pancreatic cancer is similar to findings from conventional observational studies, including the IARC Handbook Working Group(40), who reference Genkinger et al.(4) as the largest meta-analysis of body fatness on pancreatic cancer (OR for highest BMI category vs normal: 1.5, 95% CI: 1.2 to 1.8). Our results also agree with the BMI finding in a MR study using PanScan data by Carreras-Torres et al. (OR per SD increase in BMI: 1.3, 95% CI: 1.1 to 1.7)(41). Additionally, they did not change substantially in our sensitivity analyses, thus likely do not violate MR assumptions and are compatible with a causal effect.

Hip circumference, whilst potentially reflecting the observational association of general adiposity with pancreatic cancer, has not been previously implicated as a specific risk factor. Despite evidence of heterogeneity in effect-estimates when we stratified our analyses by PanScan study and sex, the direction of effect of sex- and study-specific estimates for hip circumference were the same. Thus, only the magnitude of the positive effect is uncertain for hip-circumference. The SNPs for hip

circumference show little evidence of sex-specific effects in the original GWAS(42), but consistent with findings in observational studies(43), the observed heterogeneity in this study suggests the effect of hip circumference on pancreatic cancer is stronger in females than males. Alternatively, the observed heterogeneity could reflect differences in strength of association between the IV SNPs and hip circumference between males and females; a violation of two-sample MR assumptions, casting doubt on the reliability of this result.

For the BMI and hip circumference analyses, 17 SNPs were common IVs for both phenotypes (Supplementary Table 4). Phenotype heterogeneity does not preclude causal inference, but it does undermine the ability to infer causality for particular dimensions of heterogeneous exposures, making interpretation of MR analyses more difficult(44). As a sensitivity analysis, we repeated MR analysis of these phenotypes after removing common SNPs between the hip circumference and BMI IVs(45). We obtained odds ratio estimates similar to our original estimates (BMI OR: 1.49, 95% CI: 1.17 to 1.88; hip circumference OR: 1.41, 95% CI: 1.18 to 1.69), thereby providing evidence that the original observed phenotypic associations of BMI and hip circumference with pancreatic cancer were independent; not notably driven by their shared dimension.

To our knowledge, the two metabolites ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine have not previously been associated with pancreatic cancer. There was clear heterogeneity amongst the SNPs used as instruments for these metabolites, with the associations being largely attributable to a single SNP (rs651007). The other SNPs (rs601338 and rs6151429, instrumenting ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine, respectively) showed weaker evidence of an association with pancreatic cancer. This suggests that the association of these metabolites with pancreatic cancer could reflect horizontal pleiotropy, and that the effect of rs651007 on pancreatic cancer may be mediated by some other pathway. A lookup of rs651007 in the National Human Genome Research Institute - European Bioinformatics Institute (NHGRI-EBI) GWAS Catalog revealed it to be mapped to the ABO gene; a locus that has been shown to be significantly associated with risk of pancreatic cancer genetically(17) and observationally(46). The ABO locus is associated with the serum inflammatory markers of tumour necrosis factor-alpha (TNF-α)(47) and soluble intercellular adhesion

molecule 1 (sICAM-1)(48). Inflammation has been reported to play an important role in the initiation of pancreatic tumours(49); the ABO locus may therefore influence pancreatic cancer risk by affecting systemic inflammation, thus promoting pancreatic carcinogenesis. Alternatively, these metabolites may cause pancreatic cancer, but rs601338 and rs6151429 could be subject to negative pleiotropy or not truly be associated with the metabolites, biasing our results towards the null. The limited availability of SNPs that could be used as instruments for ADpSGEGDFXAEGGGVR* and O-sulfo-L-tyrosine constrained our ability to conduct sensitivity analyses to investigate these results further.

Our results suggest evidence of an association between pancreatic cancer and genetic liability to lung adenocarcinoma. A potential explanation for this finding is sample overlap between our exposure and outcome, as we cannot reject the possibility that the PanScan control population contained individuals who were lung adenocarcinoma cases. However, Wolpin et al. report using cancer-free controls in their PanScan GWAS manuscript(50); sample overlap would therefore need to be undiagnosed lung cancer cases at the time of study. Given that the 5-year prevalence of lung cancer in the general population of Europe is 4.1%(51), we find it unlikely that there would be enough sample overlap to substantially bias our effect estimate in this instance. The association between pancreatic cancer and lung adenocarcinoma more likely reflects a shared genetic architecture with pancreatic cancer that is translated in opposing directions to affect risk in these two diseases. In either case, our finding should not be interpreted as a direct causal effect of lung cancer on pancreatic cancer (or *vice versa*). The association between these SNPs and pancreatic cancer require validation in larger GWAS and independent replication.

Smoking and type 2 diabetes, although previously reported risk factors(2, 3, 52, 53), did not show strong evidence of an association with pancreatic cancer in our analysis. Whilst the lack of association shown for smoking in our analysis could indicate that previous observational associations are biased due to confounding or reverse causation, it is also possible that our results reflect low power. The SNPs comprising the instrument for smoking (cigarettes per day) are within the CHRNA3 gene region, which is reported to proxy for smoking heaviness *amongst smokers* rather than being representative of cigarettes per day in a general population(54, 55). As such, the outcome GWAS data

would have to be restricted to current smokers to produce a meaningful effect-estimate. We couldn't stratify in this way due to the sole use of summary GWAS statistics, therefore the effect-estimate generated by our analysis is not conclusive.

Numerous meta-analyses and pooled analyses have been performed looking at the association of diabetes and pancreatic cancer, all showing that long-term diabetes is associated with a ≥50% increased risk of pancreatic cancer(56-62). Our analysis found little evidence to suggest genetic liability to type 2 diabetes has a causal effect on pancreatic cancer; a finding also reported by Carreras-Torres et al.(41).

Strengths

We appraised the association of a multitude of phenotypes with a rare cancer type in a hypothesis-free manner. Our approach features a two-sample MR design, utilising summary-level data; a particularly valuable method when the outcome of interest is rare, or when the capacity to investigate phenotypes in single studies is limited. For example, given limited power and sample size due to the cost of metabolomic platforms, many metabolites would unlikely have been investigated in relation to pancreatic cancer risk in observational studies. However, since genetic instruments for a multitude of metabolites have been obtained in previous studies with large sample sizes (63, 64), the two-sample MR framework allows the appraisal of the causal effect of the metabolome on health and disease.

Limitations

One limitation of the approach applied here is that not all possible phenotypes have genetic instruments or have not yet been curated in MR-Base. Therefore, some potentially associated phenotypes (e.g. occupational phenotypes and chronic pancreatitis) with pancreatic cancer could not be appraised.

Due to the multiple testing burden of this analysis, there was potential for false-negative findings. To remain conservative in such a broad approach, we chose to only present phenotypes that surpassed a strict Bonferroni correction in our main analysis. However, phenotypes showing weaker evidence for association (uncorrected P value <0.05) may contain some true associations and have

therefore been included in our Supplementary Materials ($p<0.05$; see **Supplementary Table 3**). On the other hand, the MR approach may identify false positive findings, particularly if there is a horizontal pleiotropic effect of a genetic instrument on the outcome, which was evident for some of the phenotypes identified here.

Given a binary outcome of pancreatic cancer, our MR models (maximum likelihood and IVW) are two-stage estimators where the second stage uses a log-linear regression model to derive an OR parameter. Estimates from such an approach will be overly precise, as uncertainty in the first-stage regression is not accounted for(65). However, this over-precision may be slight if the standard error in the first-stage coefficients is low, and can be resolved by using a maximum likelihood method(65). We provide maximum likelihood estimates in addition to IVW estimates in our MR-PheWAS analysis; these estimates are similar across our main findings, indicating that the two-stage estimator with a logistic second-stage model is still a valid test of the null hypothesis here.

By systematically evaluating the association of all available phenotypes with GWAS data in the MR-Base repository of summary genetic data, we may not have had sufficient power to detect a true causal association for every analysis conducted; particularly those proxied by low numbers of SNPs, which may infer a low phenotypic variance explained. Low numbers of SNPs to proxy a phenotype are particularly prevalent when assessing the causal association of metabolites (Kettunen et al.; Shin et al.) with pancreatic cancer; these phenotypes account for 255 of the 486 phenotypes tested, with a median 2 SNPs per metabolite. However, precise measurement of metabolites via nuclear magnetic resonance (NMR) and liquid chromatography-mass spectronomy (LC/MS) result in relatively large metabolite GWAS per-allele effect sizes and phenotypic variance explained(63, 64). The median variance explained by our metabolite phenotypes was 1.8%; at this variance explained, we had 80% power, with a median sample size of 14905, to detect an OR of 1.19 (beta of 0.17) at an alpha of 0.05.

Conclusions

Within the context of a highly aggressive cancer for which the underlying causes are poorly understood, we undertook an MR-PheWAS study which was able to suggest a causal association of a

previously identified phenotype for pancreatic cancer in observational epidemiological literature (BMI), suggest association between an anthropometric phenotype (hip circumference) with pancreatic cancer, and provide insights into some potentially novel mechanisms (metabolic factors and shared genetic architecture with lung cancer) for this disease.

References

1. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, et al. GLOBOCAN 2012 v1.1: Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 Lyon, France2014 [Available from: http://globocan.iarc.fr.
2. Wang Y, Duan H, Yang X, Guo J. Cigarette smoking and the risk of pancreatic cancer: a case-control study. Med Oncol. 2014;31(10):184.
3. Huxley R, Ansary-Moghaddam A, Berrington de González A, Barzi F, Woodward M. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. Br J Cancer. 2005;92(11):2076-83.
4. Genkinger JM, Spiegelman D, Anderson KE, Bernstein L, van den Brandt PA, Calle EE, et al. A pooled analysis of 14 cohort studies of anthropometric factors and pancreatic cancer risk. Int J Cancer. 2011;129(7):1708-17.
5. Raimondi S, Lowenfels AB, Morselli-Labate AM, Maisonneuve P, Pezzilli R. Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. Best Pract Res Clin Gastroenterol. 2010;24(3):349-58.
6. Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: unlocking the file drawer. Science (New York, NY). 2014;345(6203):1502-5.
7. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human molecular genetics. 2014;23(R1):R89-98.
8. Pierce BL, Kraft P, Zhang C. Mendelian Randomization Studies of Cancer Risk: a Literature Review. Current Epidemiology Reports. 2018:pp 1-13.
9. Yarmolinsky J, Wade KH, Richmond RC, Langdon RJ, Bull CJ, Tilling KM, et al. Causal inference in cancer epidemiology: what is the role of Mendelian randomization? bioRxiv. 2017.
10. Theodoratou E, Palmer T, Zgaga L, Farrington SM, McKeigue P, Din FVN, et al. Instrumental Variable Estimation of the Causal Effect of Plasma 25-Hydroxy-Vitamin D on Colorectal Cancer Risk: A Mendelian Randomization Analysis. PLoS ONE. 2012;7(6):e37662.
11. Hägg S, Fall T, Ploner A, Mägi R, Fischer K, Draisma HH, et al. Adiposity as a cause of cardiovascular disease: a Mendelian randomization study. International Journal of Epidemiology. 2015;44(2):578-86.
12. Pei Y, Xu Y, Niu W. Causal relevance of circulating adiponectin with cancer: a meta-analysis implementing Mendelian randomization. Tumor Biology. 2015;36(2):585-94.
13. Telomeres Mendelian Randomization C, Haycock PC, Burgess S, Nounu A, Zheng J, Okoli GN, et al. Association Between Telomere Length and Risk of Cancer and Non-Neoplastic Diseases: A Mendelian Randomization Study. JAMA Oncol. 2017;3(5):636-51.
14. Millard LAC, Davies NM, Gaunt TR, Davey Smith G, Tilling K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int J Epidemiol. 2017.
15. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. Elife. 2018;7.
16. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014;42(Database issue):D975-9.
17. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009;41(9):986-90.
18. Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. Nat Genet. 2010;42(3):224-8.
19. Childs EJ, Mocci E, Campa D, Bracci PM, Gallinger S, Goggins M, et al. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. Nat Genet. 2015;47(8):911-6.
20. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7.
21. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods. 2011;9(2):179-81.

22. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012;44(8):955-9.
23. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
24. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
25. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184-91.
26. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genet Epidemiol. 2013;37(7):658-65.
27. Burgess S, Bowden J. Integrating summarized data from multiple genetic variants in Mendelian randomization: bias and coverage properties of inverse-variance weighted methods. arXiv. 2015;arXiv:1512.04486v1.
28. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9.
29. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG, Consortium E-I. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. Eur J Epidemiol. 2015;30(7):543-52.
30. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc. 1943;54:426-82.
31. Burgess S, Labrecque JA. Mendelian randomization with a binary exposure variable: interpretation and presentation of causal estimates. Eur J Epidemiol. 2018;33(10):947-52.
32. Zheng J, Richardson T, Millard L, Hemani G, Raistrick C, Vilhjalmsson B, et al. PhenoSpD: an atlas of phenotypic correlations and a multiple testing correction for the human phenome. bioRxiv. 2017.
33. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016;32(13):1981-9.
34. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004;74(4):765-9.
35. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb). 2005;95(3):221-7.
36. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int J Epidemiol. 2015;44(2):512-25.
37. Cochran WG. The Combination of Estimates from Different Experiments. Biometrics. 1954;10(1):101-29.
38. Maisonneuve P, Lowenfels AB. Risk factors for pancreatic cancer: a summary review of meta-analytical studies. Int J Epidemiol. 2015;44(1):186-98.
39. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K, et al. Body Fatness and Cancer--Viewpoint of the IARC Working Group. N Engl J Med. 2016;375(8):794-8.
40. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K, et al. Body Fatness and Cancer--Viewpoint of the IARC Working Group. N Engl J Med. 2016;375(8):794-8.
41. Carreras-Torres R, Johansson M, Gaborieau V, Haycock PC, Wade KH, Relton CL, et al. The role of obesity, type 2 diabetes, and metabolic factors in pancreatic cancer: A Mendelian randomization study. JNCI: Journal of the National Cancer Institute. 2017;109(9).
42. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Magi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518(7538):187-96.
43. Stolzenberg-Solomon RZ, Adams K, Leitzmann M, Schairer C, Michaud DS, Hollenbeck A, et al. Adiposity, physical activity, and pancreatic cancer in the National Institutes of Health-AARP Diet and Health Cohort. Am J Epidemiol. 2008;167(5):586-97.
44. Zheng J, Baird D, Borges MC, Bowden J, Hemani G, Haycock P, et al. Recent Developments in Mendelian Randomization Studies. Curr Epidemiol Rep. 2017;4(4):330-45.

45. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. The American Journal of Clinical Nutrition. 2016.
46. Wolpin BM, Chan AT, Hartge P, Chanock SJ, Kraft P, Hunter DJ, et al. ABO blood group and the risk of pancreatic cancer. J Natl Cancer Inst. 2009;101(6):424-31.
47. Melzer D, Perry JR, Hernandez D, Corsi AM, Stevens K, Rafferty I, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008;4(5):e1000072.
48. Pare G, Chasman DI, Kellogg M, Zee RY, Rifai N, Badola S, et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. PLoS Genet. 2008;4(7):e1000118.
49. Garcea G, Dennison AR, Steward WP, Berry DP. Role of inflammation in pancreatic carcinogenesis and the implications for future therapy. Pancreatology. 2005;5(6):514-29.
50. Wolpin BM, Rizzato C, Kraft P, Kooperberg C, Petersen GM, Wang Z, et al. Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. Nat Genet. 2014;46(9):994-1000.
51. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424.
52. Fuchs CS, Colditz GA, Stampfer MJ, Giovannucci EL, Hunter DJ, Rimm EB, et al. A prospective study of cigarette smoking and the risk of pancreatic cancer. Arch Intern Med. 1996;156(19):2255-60.
53. Lu Y, García Rodríguez LA, Malgerud L, González-Pérez A, Martín-Pérez M, Lagergren J, et al. New-onset type 2 diabetes, elevated HbA1c, anti-diabetic medications, and risk of pancreatic cancer. Br J Cancer. 2015;113(11):1607-14.
54. Taylor AE, Morris RW, Fluharty ME, Bjorngaard JH, Asvold BO, Gabrielsen ME, et al. Stratification by smoking status reveals an association of CHRNA5-A3-B4 genotype with body mass index in never smokers. PLoS Genet. 2014;10(12):e1004799.
55. Lassi G, Taylor AE, Timpson NJ, Kenny PJ, Mather RJ, Eisen T, et al. The CHRNA5-A3-B4 Gene Cluster and Smoking: From Discovery to Therapeutics. Trends Neurosci. 2016;39(12):851-61.
56. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, Barzi F, Woodward M. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. Br J Cancer. 2005;92(11):2076-83.
57. Stevens RJ, Roddam AW, Beral V. Pancreatic cancer in type 1 and young-onset diabetes: systematic review and meta-analysis. Br J Cancer. 2007;96(3):507-9.
58. Ben Q, Xu M, Ning X, Liu J, Hong S, Huang W, et al. Diabetes mellitus and risk of pancreatic cancer: A meta-analysis of cohort studies. Eur J Cancer. 2011;47(13):1928-37.
59. Starup-Linde J, Karlstad O, Eriksen SA, Vestergaard P, Bronsveld HK, de Vries F, et al. CARING (CAncer Risk and INsulin analoGues): the association of diabetes mellitus and cancer risk with focus on possible determinants - a systematic review and a meta-analysis. Curr Drug Saf. 2013;8(5):296-332.
60. Li D, Tang H, Hassan MM, Holly EA, Bracci PM, Silverman DT. Diabetes and risk of pancreatic cancer: a pooled analysis of three large case-control studies. Cancer Causes Control. 2011;22(2):189-97.
61. Elena JW, Steplowski E, Yu K, Hartge P, Tobias GS, Brotzman MJ, et al. Diabetes and risk of pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium. Cancer Causes Control. 2013;24(1):13-25.
62. Bosetti C, Rosato V, Li D, Silverman D, Petersen GM, Bracci PM, et al. Diabetes, antidiabetic medications, and pancreatic cancer risk: an analysis from the International Pancreatic Cancer Case-Control Consortium. Ann Oncol. 2014;25(10):2065-72.
63. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46(6):543-50.
64. Kettunen J, Tukiainen T, Sarin AP, Ortega-Alonso A, Tikkanen E, Lyytikainen LP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. Nat Genet. 2012;44(3):269-76.
65. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Statistical Methods in Medical Research. 2015.

Tables

Exposure	# SNPs	ML OR	ML CI	P	P_{het}	R^2	F	Power	IVW OR	IVW CI
Lung adenocarcinoma	4	0.63	0.54 – 0.74	$1.68e^{-08}$	$5.47e^{-08}$	N/A	N/A	N/A	0.72	0.48 – 1.09
ADpSGEGDFXAEGGGVR*	2	1.60	1.31 – 1.95	$3.08e^{-06}$	$1.50e^{-03}$	3.59%	71.6	97.2	1.59	0.85 – 2.97
O-sulfo-L-tyrosine	2	0.58	0.46 – 0.74	$8.07e^{-06}$	$2.45e^{-04}$	1.02%	37.9	58.0	0.58	0.24 – 1.39
Hip circumference	113	1.42	1.21 – 1.67	$2.41e^{-05}$	$4.02e^{-04}$	4.46%	76.1	89.5	1.34	1.05 – 1.70
Body mass index	109	1.46	1.20 – 1.78	$1.25e^{-04}$	$1.00e^{-02}$	2.98%	91.3	81.0	1.44	1.12 – 1.86

Table 1. MR-PheWAS results passing study multiple testing threshold

Phenotypes passing multiple testing correction for the MR-PheWAS analysis. Maximum likelihood odds ratios, confidence intervals and p-values are shown for each phenotype, in addition to the number of SNPs used in the IV, a Q-test p-value for SNP heterogeneity, the variance explained, power statistics and the inverse-variance weighted odds ratio and confidence interval for each phenotype. SNP: Single-nucleotide polymorphism; ML: Maximum likelihood; OR: Odds ratio; CI: Confidence interval; Phet: P-value of heterogeneity from Q test; IVW: Inverse-variance weighted

Figure Legends

Figure 1. Smoothed line graph showing power calculations for MR analyses with X cases and Y controls (median sample size across all exposure GWAS) for genetic IVs explaining different proportions of phenotypic variance (0.1%; 0.5%; 1%; 2.5%; 5%; 10%). At our median variance explained of 3.3% with our median total sample size of 14905, we have 80% power to detect an odds ratio of 1.14 per SD change in exposure.

Figure 2. Volcano plot showing the odds ratio derived from MR analyses of 486 phenotypes against incident pancreatic cancer across the x-axis and a corresponding MR analysis p-value (-log10 scale) on the y-axis. Units are standardised - continuous traits are in standard deviation units, whereas binary traits are in log odds units. Small red points denote analyses with an unadjusted p-value < 0.05. Large red points denote analyses with a Bonferroni-adjusted p-value < 0.05

Figure 3. Forest plot of heterogeneity by PanScan study for phenotypes passing multiple testing correction. Maximum likelihood odds ratios, confidence intervals and p-values per study are given, in addition to I-squared and Q-statistics per phenotype.

Figure 4. Forest plot of heterogeneity in pancreatic cancer MR-PheWAS by sex for phenotypes passing multiple testing correction. Maximum likelihood odds ratios, confidence intervals and p-values for each sex are given, in addition to I-squared and Q-statistics per phenotype.

Figures

Figure 1. Smoothed line graph showing power calculations for MR analyses with X cases and Y controls (median sample size across all exposure GWAS) for genetic IVs explaining different proportions of phenotypic variance (0.1%; 0.5%; 1%; 2.5%; 5%; 10%). At our median variance explained of 3.3% with our median total sample size of 14905, we have 80% power to detect an odds ratio of 1.14 per SD change in exposure.

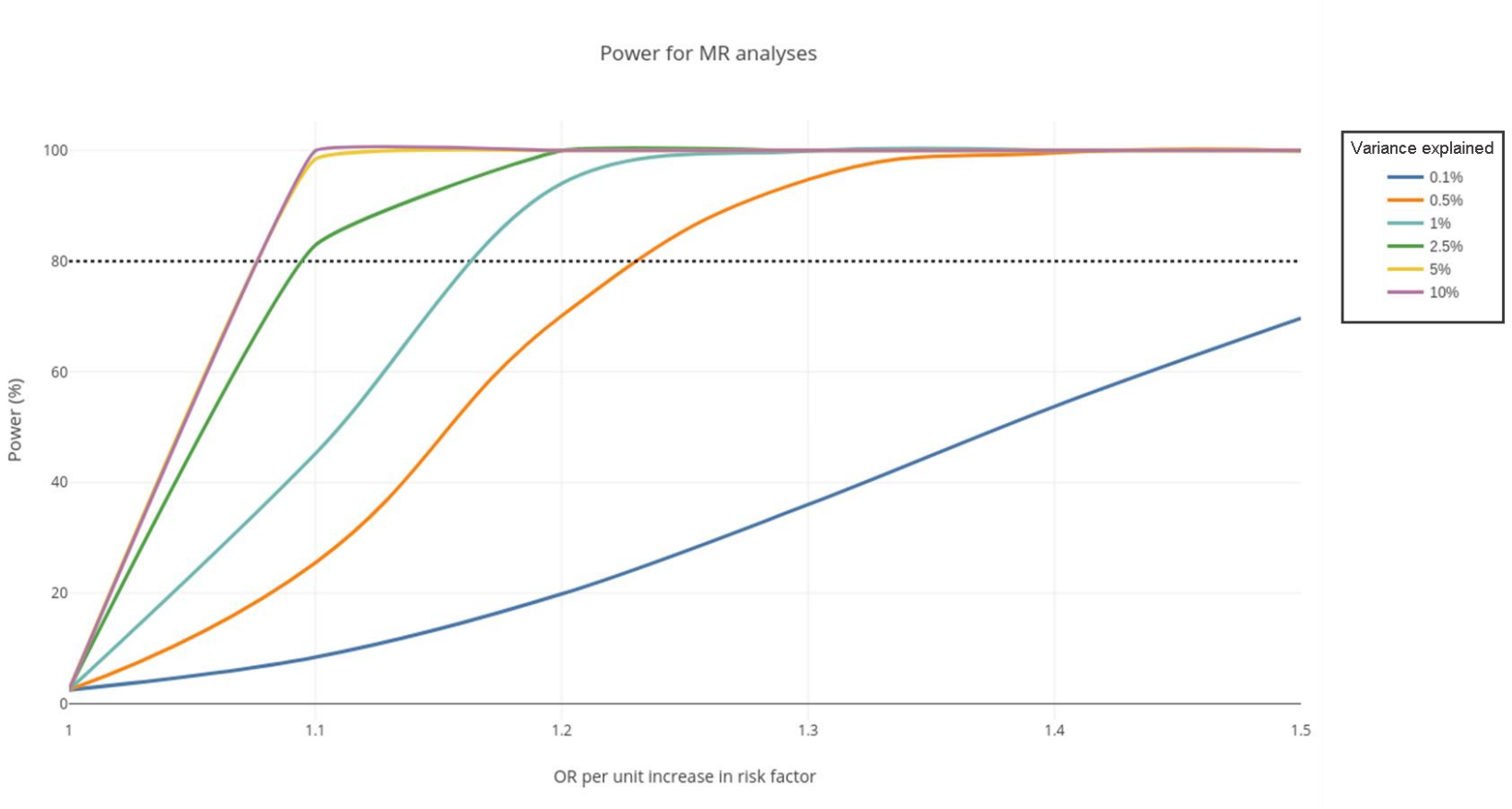

Figure 2. Volcano plot showing the odds ratio derived from MR analyses of 486 phenotypes against incident pancreatic cancer across the x-axis and a corresponding MR analysis p-value (-log10 scale) on the y-axis. Units are standardised - continuous traits are in standard deviation units, whereas binary traits are in log odds units. Small red points denote analyses with an unadjusted p-value < 0.05. Large red points denote analyses with a Bonferroni-adjusted p-value < 0.05

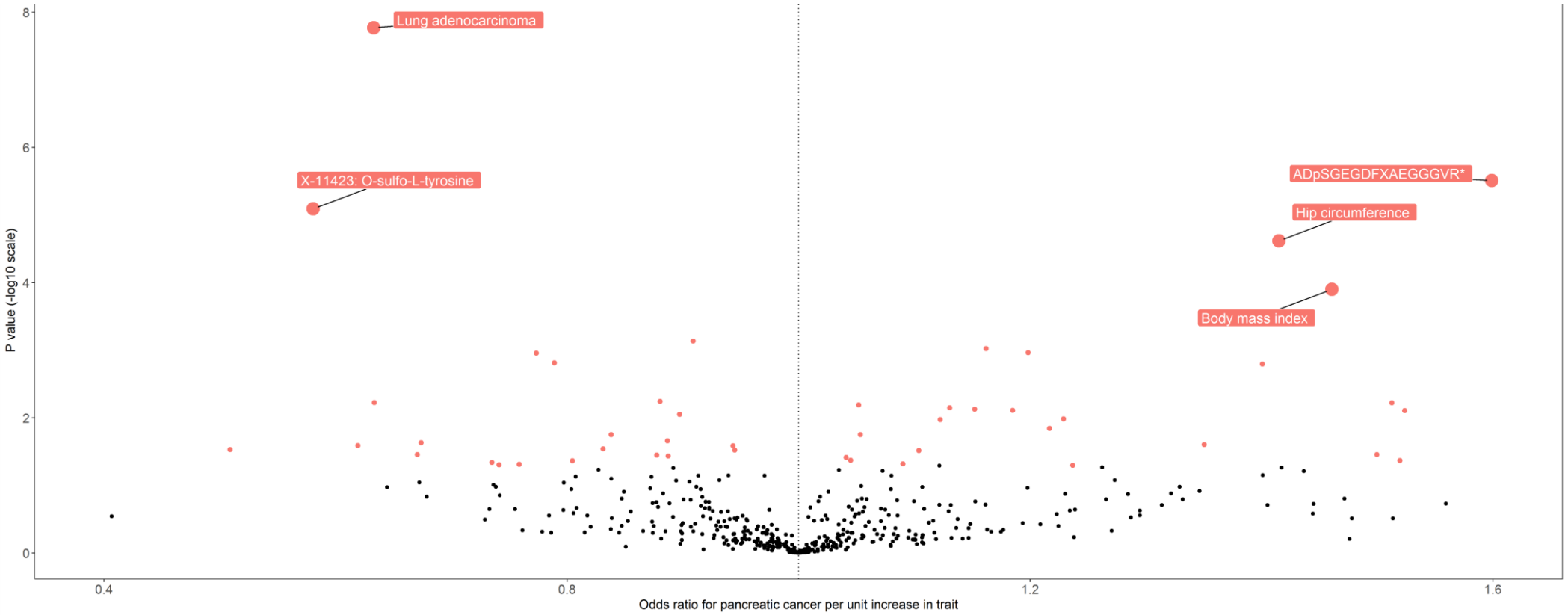

Figure 3. Forest plot of heterogeneity by PanScan study for phenotypes passing multiple testing correction. Maximum likelihood odds ratios, confidence intervals and p-values per study are given, in addition to I-squared and Q-statistics per phenotype.

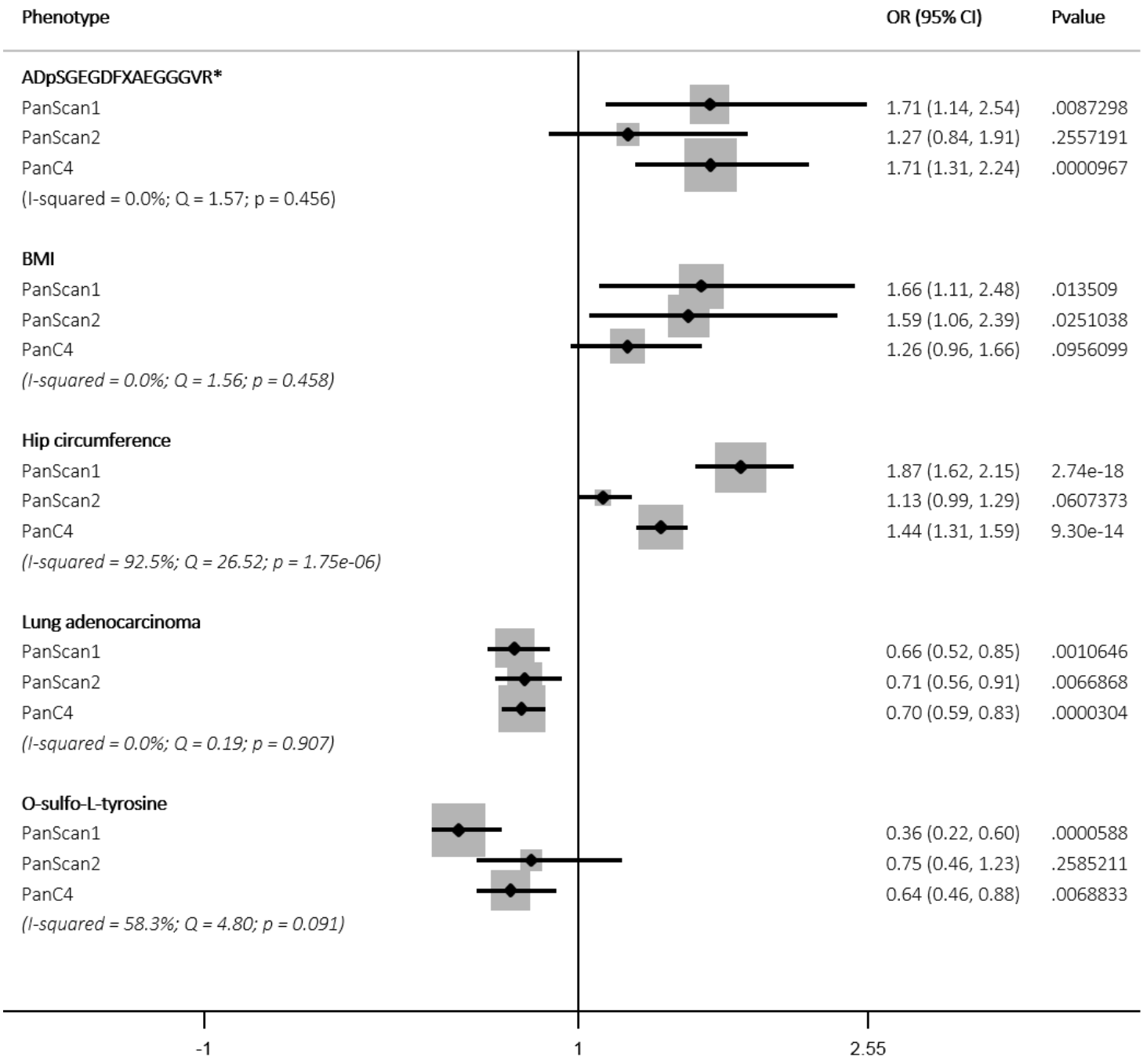

Figure 4. Forest plot of heterogeneity in pancreatic cancer MR-PheWAS by sex for phenotypes passing multiple testing correction. Maximum likelihood odds ratios, confidence intervals and p-values for each sex are given, in addition to I-squared and Q-statistics per phenotype.